A MULTIDISCIPLINARY APPROACH TO THE DEVELOPMENT OF LOW-COST HIGH-PERFORMANCE LIGHTWAVE NETWORKS

Jacek Maitan and Alex Harwit
Lockheed Missiles & Space Company, Inc., Research & Development Division,
3251 Hanover Street, Palo Alto, CA 94304-1191
e-mail: jmaitan@isi.edu

ABSTRACT

Our research focuses on high-speed distributed systems. We anticipate that our results will allow the fabrication of low-cost networks employing multi-gigabit-per-second data links for space and military applications. The recent development of high-speed low-cost photonic components and new generations of microprocessors creates an opportunity to develop advanced large-scale distributed information systems. These systems currently involve hundreds of thousands of nodes and are made up of components and communications links that may fail during operation. In order to realize these systems, research is needed into technologies that foster adaptability and scaleability. Self-organizing mechanisms are needed to integrate a working fabric of large-scale distributed systems. The challenge is to fuse theory, technology, and development methodologies to construct a cost-effective, efficient, large-scale system.

SCOPE OF THE PROBLEM

Designers of future large-scale structures for space applications must be able to solve the problems associated with access to large amounts of data distributed at various sites. The large size and distributed nature of the space applications dictates that the data be accessed with low latency and wide bandwidth. Space distributed systems of the future will need to be more reliable and require less management than existing commercial networks. Once installed, the space network must be capable of detecting, diagnosing, and recovering from both software and hardware failures. The system must be maintained though the use of a highly decentralized control structure in order to accommodate rapid changes in the system configuration and traffic patterns. Thus, the reliability requirements favor a distributed control solution. The ultimate control objective is a high degree of continuous system availability [Maitan 89a, 89b]. The number of such large-scale distributed applications will continue to grow [Chlamtac 90, Hinton 88, Nussbaum 88].

This paper is organized into three parts. In the first part, we introduce the concept of a multigrid network architecture (MNA) [Maitan 90a, 90b, 91]. In the second, we discuss the operation of MNA and suggest possible insertion points for new photonic technologies. Finally, we discuss approaches to increase the performance of MNA from 1 Gbps using state-of-the-art electronics to 50 Gbps per link using anticipated lightwave technologies. In this discussion we focus on issues associated with combining high-speed networks with low-overhead protocols and how this process affects the architecture.

ISSUES

Data networks for computer communication must offer high bandwidth (gigabits-per-second), and low latency [Young 87] and be able to handle highly variable multimedia traffic [Lidinsky 90]. Existing solutions such as high-performance parallel interface (HIPPI) or asynchronous transfer mode (ATM) were designed to address some of these needs. HIPPI is designed to handle point-to-point data transfers only. Although ATM implementations are fast, ATM requires the path to be established before transferring data. The approach works well for telephony; however, it is insufficient when a single computer broadcasts data to a large group of computers. We conclude, therefore, that systems based on the existing standards may not be able to offer effective solutions to space networking needs.

Fiber-optic networks are characterized by a small ratio between the packet transmission time and the packet propagation time through the network. The latter is usually larger because it includes the time required to route and to resolve contention at nodes. In fiber-optic networks, this is the major source of propagation delays. An increase in the transceiver speed to decrease the packet transmission time will not increase the actual network throughput.

Increased transmission speeds change the data communication strategy. Bandwidth is cheap and computation is expensive. In traditional connection-oriented systems, the traffic is controlled by a simple store-and-forward algorithm. This strategy requires careful management of the buffering and protocol processes [Clark 89], and is effective only when the computers are much faster than the networks. In state-of-the-art systems, i.e., a 32-bit computer operating at 100 million instructions per second (MIPS), the total data flow is 3.2 Gbps. This is the same order of magnitude as the 1 Gbps available in a communication link. Thus, the ability to handle high-speed traffic by simple store-and-forward algorithms using general purpose computers has disappeared.

The key solution to all of these problems is an effective control strategy. Today's careful bit stuffing, used to increase bandwidth utilization, results in complex protocols, which must be replaced with bandwidth-effective data transmission protocols with lower processing overhead.

MULTIGRID NETWORK ARCHITECTURE

Our research has focused on providing a network which is flexible, scaleable, simple to control, and very effective in its operation. To address these objectives, we simplify the medium access control (MAC) protocol and eliminate as much of the protocol processing overhead as possible. This begins with the elimination of the store-and-forward transport algorithms. The result is a novel bufferless control structure for very high-performance packet-switching networks. A full multigrid network architecture (MNA) implementation features both circuit and packet switching [Maitan 90c]. In this paper, the discussion is limited to packet switching only.

A network switch, also referred to as a node, is shown in Fig. 1. Here, a data packet must bid for an output link by providing an ordered list of links which can be used. The router must simply assign an output link for each packet, based on all the bids. Conflicting bids for the same output port are nondeterministically solved. Thus, the whole network is treated as both a buffer and a distributed data processing structure that is capable of routing packets toward their desired destination. The bidding process is strictly local and solves the problem of central control, which would otherwise be nearly impossible to achieve. MNA is a connectionless network and packets from all routing nodes and hosts are treated equally when competing for resources. No restriction is placed on the topology or size of the network.

Most existing networks use dynamic routing, in which priorities are computed based on traffic estimates. Thus, if not controlled, the asynchronous nature of dynamic routing table updates may result in packet looping. In MNA, the tables are statically computed and permanent loops can be detected and eliminated. In an MNA routing table, a failed link is simply marked as a busy link. This routing mechanism substantially simplifies the complexity of controlling the node and, as we have demonstrated, leads to a very simple hardware system implementation.

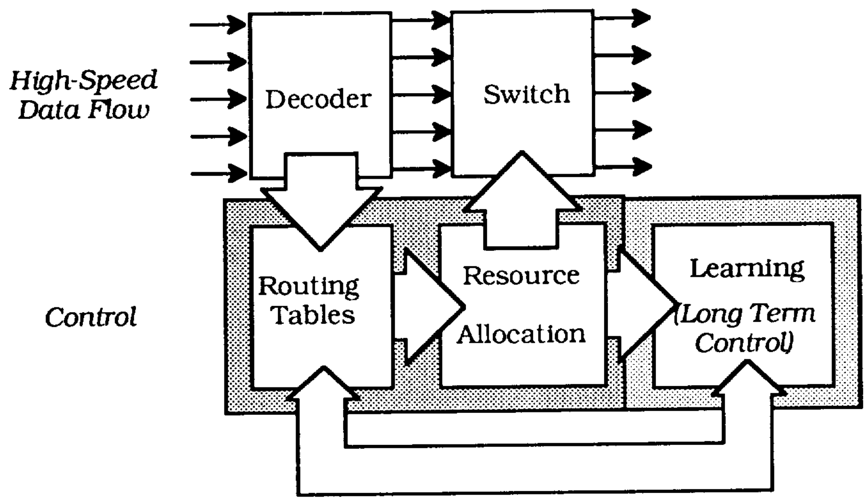

Fig. 1 A simplified block diagram of a router for a packet switch

In a well-connected network of arbitrary topology, there is usually more than one output link suitable for routing, Fig. 2. Access to a particular link is given using a stochastic resource-allocation policy and priority is given to packets that are already in traffic. Thus, in highly congested traffic, extra packet bursts diffuse to neighboring nodes, so that, the network as a whole is able to store extra traffic. The failure of a link may cause an imbalance in the system. For example, if the number of incoming packets at a node exceeds the number of outgoing links, packets may be lost. Higher level protocols must be used to control the recovery of missing data as discussed later in this paper.

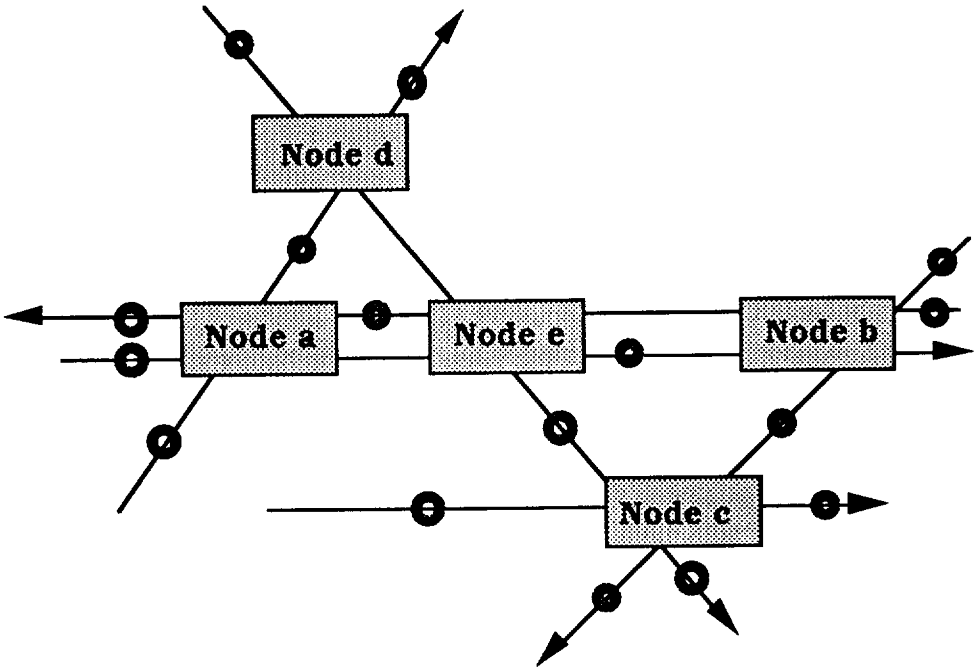

Fig. 2 Interaction between nodes in MNA

We have built prototype hardware and performed computer simulations [Gburzynski 90] to show that a distributed control scheme of this type applied to such networks is very effective for a wide range of parameters. The lengths of the routing paths are on average independent of load and the physical lengths of the links.

MNA PROTOCOL

All data is transferred as fixed-size packets consisting of a header and a payload. The header contains control and error-detection information in addition to the destination address. Packets can be sent as single entities or in groups. They are generated by the hosts which are connected to the routing nodes. To be transferred, a packet must first gain access to the transport fabric.

On arrival at a node, a packet submits a bid to the resource allocator for an output port. If all desired ports are busy, an arbitrary free port is assigned to keep a packet in transition. A packet is lost only when no free output ports are available. Once access to a port is granted, it is used to transmit the entire packet. MNA does not use local storage; instead, it uses the whole network as a buffer.

In the case of substantial system failures combined with a heavy load, packet losses are unavoidable and must be handled by an appropriate transport protocol. Thus, under catastrophic conditions, MNA converges to the traditional connection-oriented network with circuit reservation. However, unlike traditional fault-tolerant networks, all resources are used; none are kept simply as a reserve. The soft failure feature is built into MNA and is not added as an afterthought.

In all-optical networks, one can split the signal and process only the header. This lack of intermediate buffering simplifies the construction of an all-optical switching network, as discussed below.

MNA SYSTEM INTEGRATION

In this section, we outline the approach we are taking in the hardware prototype that is currently under construction. We also discuss how MNA scales up with new high-speed technologies.

The packet-switching circuitry is currently being constructed and tested in an all-electrical implementation. It is composed of off-the-shelf CMOS devices and gate arrays. To date, we have constructed and tested circuits to process an estimated 10^7 resource allocation bids per second. In the future, the network is envisioned to utilize an all optical switching fabric.

In a photonic network, the nodes are connected to each other by optical fibers. Figure 3 shows a block diagram of an 8 x 8 high-speed network switch. This switch consists of two 8-input fiber couplers and routing circuitry to route incoming data to the appropriate output. Data packets are assumed to arrive on single-mode 1550-nm optical fibers, formatted in 500-bit packets at 50 Gbps. The total packet length is 10 ns.

Optical data enters the switch through the input fiber coupler. In order to correctly route an incoming packet, a small amount of optical energy is removed from the incoming signal to form the control path. Energy in the control path is converted to an electrical signal from which a destination address is extracted. The destination address is then used as the input for a routing circuit that performs resource allocation and finally configures the switch. The remainder of the signal flows into a temporary buffer which consists of a small length of Er^+-doped fiber that acts both as a delay line and wide-bandwidth amplifier to boost the optical signal. After amplification, the output of the buffer goes to an optical switch which routes the data to the appropriate output on the output fiber coupler.

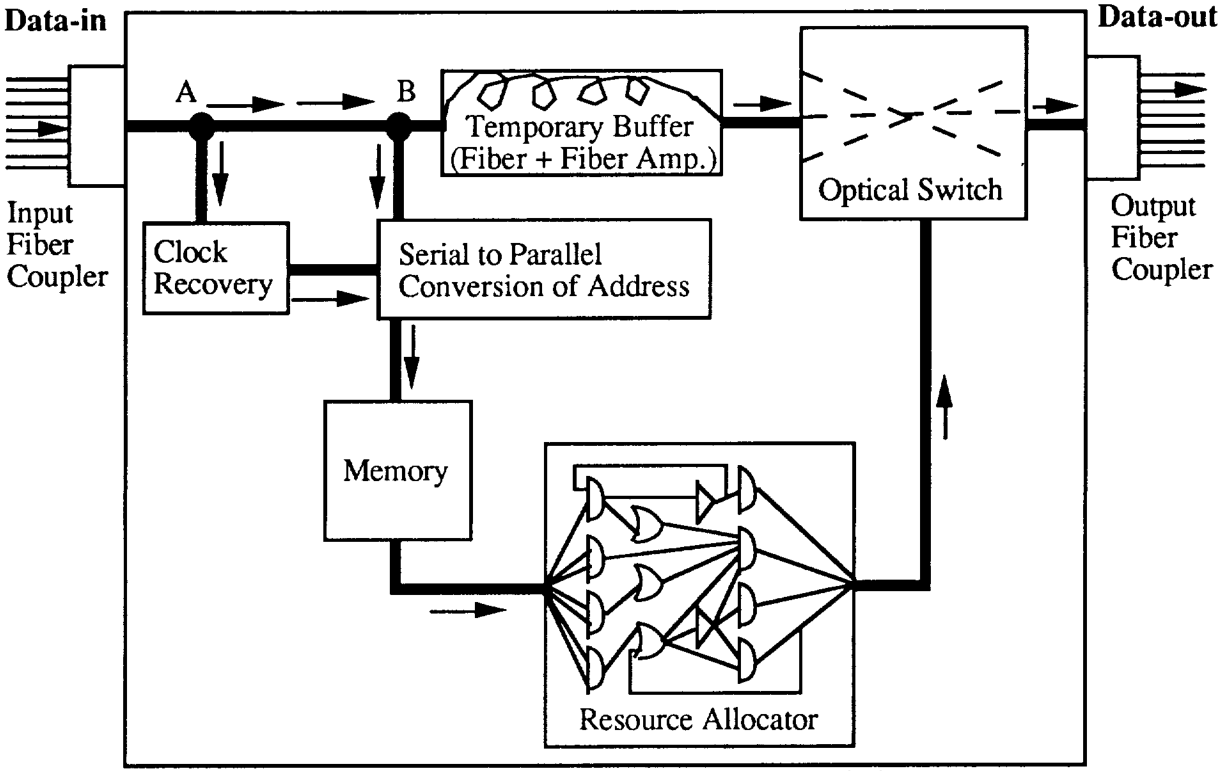

Fig. 3 An 8 x 8 high speed network switch

In the control path, the switch must recover the clock and extract the header of the packet. Specifically, a small amount of optical energy is tapped off at point "A" for a clock-recovery circuit [Swartz, 88]. The function of this circuit is to align the clock of the node that generated the data with the local clock of the switch. There is one clock-recovery circuit per input line. At point "B", a second optical tap removes additional signal and sends it to a serial-to-parallel converter. With the clock timing information from the clock-recovery circuit, this circuit extracts the destination address from the packet header and converts it to a parallel format. The destination address is then used as an address for a routing table that is simply a high-speed memory. Each memory location is associated with a destination and contains an ordered list of desired output channels to be used as bids. There is one serial-to-parallel converter and one memory for each input line. The bids from each memory are all processed by a single resource-allocator unit. The resource-allocator unit is composed of a large sequence of discrete logic gates. The output of the resource allocator configures an optical switch. To properly control the switch, the spacing between packets must be greater than the sum of the optical switch switching and settling times.

Several constraints exist on each of the circuits in the network switch. The fiber coupler, for example, must be able to handle at least eight single-mode 1550-nm fibers with a low insertion loss. It will most likely be composed of a V-groove technology in which the fibers are pigtailed into the substrate containing the electronic processing circuitry. Since each packet is about 10 ns long, the fiber line should delay the signal by about 20 ns, during which time the circuit could process the packet destination address and set the optical switch. The clock extraction circuit must be able to extract the 50 Gbps timing in about 3 ns. The serial-to-parallel converter must also be able to extract the packet destination address in about 3 ns. For an 8 x 8 switch utilizing a bid format consisting of three options, the memory size would be equal to: [number of nodes in network] x [3 x $\log_2$(number of nodes)]. To balance the flow in the data path, the memory is required to have an access time of about 4 ns, and such parts are available today. The resource allocator is composed of discrete "AND", "OR", and "NOR" gates, etc., and would be required to switch in about 4 ns. The low-loss optical switch may either be an integrated optoelectronic polymer switch with active rail taps [Van Eck 91] or a multiple-quantum-well modulator [Komatsu, 90]. It is required to switch in about 4 ns.

The key point in the MNA approach is the reduction of the packet propagation time by careful management of routing information combined with the application of a new class of evaluators to resolve synchronization and resource-allocation problems. This is especially important in local area networks (LANs) where one must also be able to manage latency. Instead of optimizing protocol layers one at the time, we have attempted to consider the interaction of mechanisms associated with several protocol layers simultaneously.

SUMMARY

In this paper, we have discussed a completely distributed packet-routing architecture called multigrid network architecture (MNA), for building low-cost high-speed networks. To prove the feasibility of such networks, we have prototyped a resource allocator which has an estimated performance of 8 x 10^6 packets/s for an 8 x 8 packet switch. Currently, we are completing a hardware prototype of an 8 x 8 pizza-box-sized packet switch capable of handling 1 Gbps traffic at each port. Furthermore, the networks are capable of transferring data at up to 50 Gbps per line and can be controlled using MNA distributed-control algorithms.

MNA is an architecture that has been designed to scale with evolving technologies. It is also an attempt to simplify and integrate an implementation of a multilayer protocol stack. This work is an attempt to identify an approach leading to the cost-effective use of high-speed networks in application-oriented distributed systems. Preliminary results are encouraging and indicate that such networks can be controlled using simple algorithms implemented in low-volume switches that can be built using existing technologies.

ACKNOWLEDGMENTS

We would like to thank L. Walichiewicz and D. Robertson for useful discussions. This work was supported in part by Lockheed Internal Research funding and NASA contract NAS2-13223.

REFERENCES

[Chlamtac 90] Chlamtac, I., and Franta, W.R., “Rationale, Directions, and Issues Surrounding High-Speed Networks,” *IEEE Proceedings,* Vol. 78, No. 1, pp. 94-120, January 1990.

[Clark 89] Clark, D.D., van Jacobson, D., Romkey, J., and Salwen, H., “An Analysis of TCP Processing Overhead,” *IEEE Communications Magazine,* Vol. 27, No. 6, pp. 23-36, June, 1989.

[Gburzynski 90] Gburzynski, P., and Rudnicki, P., *The LANSF Protocol Modeling Environment, version 2.0,* Department of Computer Science, University of Alberta, Edmonton, Alberta, Canada, 1990.

[Hinton 88] H. S. Hinton. “Architectural Considerations for Photonic Switching Networks”. *IEEE Journal on Selected Areas in Communications,* Vol. 6, No. 7, pp. 1209-1226 August 1988.

[Komatsu, 90] Komatsu, K., Sugimoto, M., Ajisawa, A., Hamamoto, K., Kohga, Y., and Suzuki, A., "4 X 4 GaAs/AlGaAs Optical Matrix Switches with Electro-Optic Guided-Wave Directional Couplers", in *Photonic Switching II*, editors Tada, K and Hinton, H. S. , Springer Verlag, Berlin, 67-71, 1990.

[Lidinsky 90] Lidinsky, W. P., Data Communications Needs, *IEEE Network,* Vol. 2, No. 4, pp.28-33, March 1990.

[Maitan 89a] Maitan, J., “Intelligent Distributed and Networking Systems”, *Proceedings of IJCAI workshop on The Future of Research in Artificial Intelligence,* IFIP's Artificial Intelligence Specialists Group sponsored by IFIP and IJCAI, 1989.

[Maitan 89b] Maitan, J., Ras, Z., and Zemankova, M., “Querry Handling and Learning in a Distributed Intelligent System”, *Proceedings of the Fourth International Symposium on Methodologies for Intelligent Systems,* Charlotte, North Carolina, 1989.

[Maitan 90a] Maitan, J., Walichiewicz, L., and Wealand, B., "Communication Technology Requirements for the Next Generation of Ground Systems," Final Report, May 1990.

[Maitan 90b] Maitan, J., Ras, Z.W., "OPERA: Opto-Electronic Reconfigurable Architecture for Multiprocessor Systems, Mathematical Foundations”, *Proceedings of the Fifth International Symposium on Methodologies for Intelligent Systems,* Charlotte, North Carolina, 1990.

[Maitan 90c] Maitan, J., Walichiewicz, L., and Wealand, B., "Integrated Communication and Information Fabric for Space Applications”, *AIAA/NASA Second International Symposium on Space Information Systems,* September, 17-19, 1990

[Maitan 91] Maitan, J.: "A Flow Control Mechanism for Distributed Systems,"› *Proceedings of the SPIE Conference on Target Recognition,* Orlando Florida, 1991.

[Nussbaum 88] Nussbaum, E. “Communication Networks Needs and Technologies - A Place for Photonic Switch?” *IEEE Journal on Selected Areas in Communications,* Vol. 6, No. 7, pp. 1036-1044, August 1988.

[Swartz, 88] Swartz, Robert G., "High Performance Integrated Circuits for Lightwave Systems", in *Optical Fiber Telecommunications II,* editors Miller, Stewarte and Kaminow, Ivan P., Academic Press, Boston, 759-780, 1988.

[Van Eck 91] Van Eck, T. E., Ticknor, A. J. , Lytel, R. S. , and Lipscomb, G. F., “Complementary Optical Tap Fabricated in an Electro-optic Polymer Waveguide”, *Appl. Phys. Lett.* Vol 58, pp. 1588-1590, 1991.

[Young 87] Young, M., Tevanian, A., Rashid, R., Golub, D., *The Duality of Memory and Communication in the Implementation of a Multiprocessor Operating System,* unpublished material, Department of Computer Science, Carnegie-Mellon University, 1987.